Oregon Health & Science University
3181 S.W. Sam Jackson Park Rd.
Portland, Oregon 97239-3098

Final Report

CSSG: Learning within NLP pipelines for scalable data mining and information extraction

August 7, 2011

Reporting Period: **05/07/2009 - 05/06/2011**

Prepared For:
Defense Advanced Research Projects Agency (DARPA)
3701 North Fairfax Drive
Arlington, VA 22203-1714

Under Contract Number:
HR0011-09-1-0041

Submitted by:
Brian Roark, Ph.D., Principle Investigator
(503) 748-1752, email: roarkbr@gmail.com

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
08/07/2011	Final Report	05/07/2009 - 05/06/2011

4. TITLE AND SUBTITLE
CSSG: Learning within NLP pipelines
for scalable data mining and information extraction

5a. CONTRACT NUMBER
HR0011-09-1-0041

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Brian Roark

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Oregon Health & Science University
3181 S.W. Sam Jackson Park Rd.
Portland, Oregon 97239-3098

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency
3701 North Fairfax Drive
Arlington, VA 22203-1714

10. SPONSOR/MONITOR'S ACRONYM(S)
DARPA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This document is the Final Report of the Learning within NLP pipelines for scalable data mining and information extraction project funded by DARPA contract number HR0011-09-1-0041. This final report discusses the successful completion of the program's objectives.

15. SUBJECT TERMS
Natural Language Processing; context-free syntactic parsing; finite-state methods; pipeline systems

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Brian Roark
Unclassified	Unclassified	Unclassified	UU	4	19b. TELEPHONE NUMBER *(Include area code)* 503 748-1752

DARPA HR0011-09-1-0041
CSSG: Learning within NLP pipelines for scalable data mining and information extraction

PI, Brian Roark, Oregon Health & Science University

Final Report
Phase 2 Computer Science Study Panel

Period of Performance: The period of performance for this report is 05/07/2009 through 05/06/2011.

I. Results

Programmatic

Attendance/participation in sessions

As planned, the PI attended both sessions for the phase 2 Computer Science Study Panel (CSSP). During the first three-day session, which fell in the first quarter of the first year of the project, each participant gave short research talks about the topics of their CSSP projects, and the group visited several interesting groups, both for support of ongoing research (IDA library) and potential partners for new research (Library of Congress, Federal Reserve Board and US Army Communications Electronics Command, CECOM). The PI gave a brief description of his research project to the panel participants at IDA. During the second three-day session, which fell in the second quarter of the first year of the program, each participant again gave a short research talk about the topics of their CSSP projects, and the group had the chance to interact with members from other CSSP classes. The group also visited the office of the Director of National Intelligence, where they heard several very interesting briefings. On the last day of that session, he gave a half hour description of his research project to the panel participants at IDA.

Project/research focus

This phase 2 CSSG project focused on two specific applications that are important for national security: rich information extraction from text and term detection in speech. The project specifically focused on Chinese and English text processing. These applications are instances of pipeline systems for text and speech processing, and the project pursued novel approaches to pipelines. These included: (i) new approaches for machine learning within these pipeline systems, which incorporate penalties for both excessive later stage search and cascading errors; (ii) innovations in context-free processing algorithms to achieve significant worst-case and typical-case complexity improvements, using both cascaded constraints and fast intersection techniques; and (iii) general software libraries and tools in support of novel algorithms and data structures for use within pipeline systems for text and speech processing.

Expected impact of program participation on research

This program provided important support for the PI's research program, both in terms of direct PI time allocated to these projects, but also for student and collaborator funding. That the PI was

able to direct so much time and energy to the topics in this program means that the topics have advanced dramatically during the course of the project, and they now form a major focus of his overall research program. The work accomplished during this program has been recognized through a best paper award (see details below) and a number of publications have been produced in top tier conferences (journal papers to follow, see below for more details). This work is seen in the community to be an important contribution to the literature on the topic, and will likely lead to further funding in the future. One student who has been supported on this grant is preparing to finish his dissertation. Overall, this has been a very high impact program for the PI's work.

Technical

Potential DoD research applications

The PI's research area is natural language processing, and the applications that he works on deal with speech or text data. Over the two CSSP phases, the group has been presented with many potential applications for NLP in processing speech or text. In all cases, the key questions that are arising have to do with building algorithms that can scale up to very large collections, e.g., web scale applications. The PI is very interested in applications that attempt to extract rich information (such as representations of events) from large collections of heterogeneous data. One possible application, presented at SOCOM during the first phase, involved extraction of information to feed into predictive models. The scale of text for that effort was VERY large, which should push the research to develop approaches that are more efficient than the current state-of-the-art in structured processing. These sorts of problems are very common for both speech processing and text processing applications of importance to national security, hence of interest to groups within the DoD.

Potential transition partners

While the PI's follow up activities failed to yield a viable phase three project within this program, he continues to have contact with individuals inside of the DoD, who are very supportive of his work. He expects to find continuing funding by DoD partners at some point in the future.

Funding:

All allocated funds were spent during the period of the award. There are no issues relating to actual versus budgeted amounts.

II. Comparison of actual accomplishments with the goals and objectives established for the grant, the findings of the investigator or both.

As stated above, the goals established for the grant involved (i) new approaches for machine learning within pipeline systems, which incorporate penalties for both excessive later stage search and cascading errors; (ii) innovations in context-free processing algorithms to achieve significant worst-case and typical-case complexity improvements, using both cascaded constraints and fast intersection techniques; and (iii) general software libraries and tools in support of novel algorithms and data structures for use within pipeline systems for text and speech processing. The goals were met in all three areas.

For part (i) above, which forms the core topic of the PhD thesis work of Nathan Bodenstab, who is planning to defend this coming academic year, a publication [1] in the Annual Meeting of the Association for Computational Linguistics in 2011 did exactly what was described in the goal, by learning how to search through alternative context-free parsing structures very efficiently. Nate, in collaboration with the PI, another graduate student Aaron Dunlop, and Keith Hall from Google Research, designed an algorithm to learn to balance the efficiency of heuristic search with the accuracy of the resulting inference, to achieve substantial parsing speedups relative to the state of the art, with no loss in accuracy (and sometimes a gain). The remainder of his dissertation will examine a couple of additional alternative approaches to this problem. Overall, this work fulfilled the promise of the original ideas presented in this proposal, and that work was published this year.

For part (ii) above, two threads of now completed work were targeted at achieving complexity improvements in context-free parsing pipelines. In the first, recently graduated PhD student Kristy Hollingshead (now a post-doctoral fellow at the University of Maryland in College Park) collaborated with the PI and graduate student Nate Bodenstab on methods for using a finite-state tagger to achieve complexity improvements in context-free parsing pipelines. First, in 2009, Roark and Hollingshead [2] presented a method to guarantee linear complexity of context-free parsing pipelines given finite-state annotations, as well as related methods that provide large typical case speedups. Recent follow up work [3] showed that additional finite-state tagged constraints can yield further typical case speedups. Both of these results will be covered in a journal publication in preparation. Second, graduate student Aaron Dunlop, in collaboration with the PI and graduate student Nate Bodenstab, devised a novel factorization of the inner-loop of context-free syntactic parsing, yielding a sparse matrix operation that achieved orders-of-magnitude speedups over other techniques. That work was used in [1], and is covered in a publication currently in submission, as well as an earlier technical report [4].

All of the parsing work above has been built into an open source context-free syntactic parser, known as BUBS (http://code.google.com/p/bubs-parser). In addition to this open source library, the PI has been collaborating with co-PI's Richard Sproat and Izhad Shafran on some novel finite-state data structures for encoding statistical sequence models, which will be released as part of the open source OpenGrm library (opengrm.org). In a recent publication [5], we showed that exact off-line backoff models could be represented using a variant of the lexicographic semiring. (The paper won best short paper at ACL 2011.) The key benefit of this approach is that it allows the models to be combined off-line with other weighted finite-state transducers and optimized prior to use in the particular application. This work has led to some other projects (publications under review) in the same vein, that exploit rich semiring data structures to straightforwardly encode finite-state syntactic models of high utility for spoken language processing. Beyond this, the PI has contributed core functionality to the above-cited OpenGrm library, which is nearly completely released at the above URL.

In sum, the research direction of the PI's lab group over the past year remained very close to the original goals and objectives of the grant, and substantial progress was made in all three areas. All of the areas continue to yield fruitful research directions, and hence the PI will continue to work in these areas.

III. Reasons why established goals were not met, if appropriate

Not applicable, established goals were met.

IV. Other pertinent information

References:

[1] N. Bodenstab, A. Dunlop, K. Hall and B. Roark. 2011. Beam-Width Prediction for Efficient CYK Parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 440-449.

[2] B. Roark and K. Hollingshead. 2009. Linear complexity context-free parsing pipelines via chart constraints. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL),* pp. 647-655.

[3] N. Bodenstab, K. Hollingshead and B. Roark. 2011. Unary Constraints for Context-Free Parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL),* pp. 676-681.

[4] A. Dunlop, N. Bodenstab and B. Roark. 2010. Reducing the grammar constant: an analysis of CYK parsing efficiency. *Technical Report CSLU-2010-02*, Center for Spoken Language Processing, Oregon Health & Science University

[5] B. Roark, R. Sproat and I. Shafran. 2011. Lexicographic Semirings for Exact Automata Encoding of Sequence Models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1-5. (Best short paper award.)